Modèle, Calibration et Leviers d'Amélioration

1 Modèle sélectionné, architecture et techniques utilisées

1.1 Architecture retenue

Classifieur principal : `CatBoostClassifier`
Tâche : prédire *Accident* vs *Non-accident*.

Atouts CatBoost (données tabulaires hétérogènes).

- Gestion native des variables catégorielles (encodage ordonné, one-hot implicite).
- Tolérance aux valeurs manquantes.
- Modélisation d'interactions non linéaires.
- Gestion du déséquilibre via `class_weights` / `scale_pos_weight`.

Calibration des probabilités.

- Méthode : régression isotone (`IsotonicRegression`).
- Objectif : transformer le score en une probabilité fiable (Brier faible, ECE $\to 0$).
- Règle : la calibration est apprise sur un *holdout* du train (jamais sur le test) afin d'éviter toute fuite.

Voie secondaire (sanity check).

- `CatBoostRegressor` calibré isotone pour produire une probabilité alternative.
- Les deux voies doivent converger (cohérence des risques estimés).

1.2 Prétraitement et anti-leaks

Prévention des fuites.

- Suppression des colonnes *fuyardes* (p. ex. indicateurs post-événement, colonnes temporaires d'exploration).
- Détection et retrait des colonnes fortement corrélées à la cible : $|\rho| \geq 0{,}98$.
- Suppression des colonnes quasi égales à la cible ($\geq$ 99,9% d'égalité).

Catégorielles et modalités rares.

- Variables de type `CSP_`, `Apporteur_`, `Carrosserie_`, `Region_`, `Usage_`, `Alimentation_`, etc.
- Regroupement des modalités très rares en `__RARE__` pour stabiliser l'apprentissage.

Dates → dérivées. À partir de Date_naissance, Date_obtention_permis, Date_debut_comerc, Date_fin_comerc :

- Year, Month, DayOfWeek, Age_in_days, ancienneté permis.

Numériques → dérivées.

- Transformations: log1p, z-score, termes au carré.
- Ratios: p. ex. Puissance_vehicule / Groupe_vehicule.
- Parts relatives (*shares*) par groupe pertinent.

Split. Partition **80/20** (train/validation ou train/test selon le protocole).

1.3 Variables clés

CatBoost tire fortement parti des familles suivantes :

- **Âge/expérience** : Date_obtention_permis, Age_obtention_permis (continu et binnings quantiles).
- **Véhicule** : Puissance_vehicule, Carrosserie_, Classe_vehicule_, Groupe_vehicule.
- **Usage** : Usage_ (*Professionnel*, *Tournées*, etc.).
- **Zone** : Region_ (densité, urbanité implicite).
- **Sociodémo/commercial** : CSP_, Situation_maritale_, Apporteur_ et croisements (Apporteur × CSP, Apporteur × Situation_maritale).

1.4 Apprentissage, métriques et décision métier

Objectifs et critères.

- En entraînement : **LogLoss**.
- En sélection : **PR-AUC** prioritaire (classe *accident* rare).
- En exploitation : **Brier** et **ECE** pour la calibration (probabilités = risques).

Métrique	**Valeur**
Accuracy@*thr*	0.88249
BalancedAccuracy@*thr*	0.9216342898536058
Brier	0.04352415031864684
ECE	0.006470668724492427
F1@*thr*	0.7477947331144164
LogLoss	0.1371007686907344
PR_AUC	0.9140265005946538
ROC_AUC	0.979717151782668
expected_cost	14577
FN	314
FP	11437
TN	70828
TP	17421

Rapport de métriques (Comet / MLOps).

Paramètre	**Valeur**
bootstrap_type	MVS
border_count	128
depth	6
eval_metric	AUC
iterations	400
l2_leaf_reg	6.0
leaf_estimation_iterations	1
learning_rate	0.1
loss_function	Logloss
od_type	Iter
od_wait	50
random_seed	42
sampling_frequency	PerTreeLevel
scale_pos_weight	4.638647288515485

Hyperparamètres principaux (classifieur).

2 Pourquoi la courbe Précision–Rappel et comment elle pilote la décision

2.1 Pourquoi PR plutôt que ROC ?

Sur données déséquilibrées, la courbe ROC peut sembler élevée tout en manquant de nombreux positifs. La courbe PR met l'accent sur :

- la **précision** (éviter d'alerter des non-accidents) ;
- le **rappel** (capturer un maximum d'accidents).

Ici, PR-AUC = 0.914, indiquant une précision élevée même à rappel moyen/élevé, ce qui est propice à des politiques de prévention/tarification.

2.2 Seuil métier par coût attendu

On minimise le coût attendu en fonction du seuil t. Soit

$$\mathbb{E}[\mathrm{Cost}(t)] \;=\; C_{\mathrm{FN}} \cdot \mathrm{FN}(t) \;+\; C_{\mathrm{FP}} \cdot \mathrm{FP}(t),$$

avec $C_{\mathrm{FN}} \gg C_{\mathrm{FP}}$ (p. ex. $C_{\mathrm{FN}} \in [10, 15]$ et $C_{\mathrm{FP}} = 1$). Le seuil retenu $t^{\star} = 0.102$ est choisi :

- au minimum de la courbe de coût (ou au milieu d'un plateau à $\pm\varepsilon$) ;
- cohérent avec l'asymétrie des coûts ;
- garantissant un compromis précision/rappel qui minimise la perte économique.

2.3 Calibration et interprétabilité

La régression isotone ajuste les scores $\hat{s}$ en probabilités calibrées $\hat{p}$ telles que

$$\Pr(Y{=}1 \mid \hat{p} \in [a, b]) \approx \hat{p}.$$

La qualité est mesurée par un **Brier** faible et une **ECE** proche de 0, assurant qu'une prédiction $\hat{p} = 0.30$ correspond bien à $\sim 30\%$ d'événements en moyenne. C'est critique pour le pricing, le bonus–malus et les décisions explicables.

Formules utiles.

$$\text{Brier} \;=\; \frac{1}{n}\sum_{i=1}^{n}\left(\hat{p}_i - y_i\right)^2, \qquad \text{ECE} \;\approx\; \sum_{b=1}^{B}\frac{|S_b|}{n}\,\big|\text{acc}(S_b)-\text{conf}(S_b)\big|.$$

3 Trois leviers d'amélioration

3.1 Feature engineering avancé

- Conserver les versions *continues* en plus des binnings (p. ex. `Age_obtention_permis`, `Puissance_vehicule`, `Groupe_vehicule`) ; CatBoost place des coupures optimales.
- Interactions ciblées (*cross-features*) : `Usage` × `Carrosserie`, `Puissance` × `Usage`, `Region` × `Usage`, `Apporteur` × `CSP`.
- Ancienneté permis : Âge_permis = date_réf − `Date_obtention_permis`.
- TE KFold ordonné pour les hautes cardinalités (`Code_vehicule`, etc.) sans fuite (jamais la même ligne en fit/transform).
- Regroupement éclairé des modalités rares selon règles métier.

3.2 Validation & optimisation orientées stabilité/coût

- Validation temporelle stricte (*out-of-time*, fenêtres glissantes) et mesure de la variance de AUC, PR-AUC, Brier, Cost.
- Seuil par plateau de coût : choisir t tel que

  $$\text{Cost}(t) \le (1+\varepsilon)\cdot \min_u \text{Cost}(u) \quad (\varepsilon \approx 0{,}1\%),$$

  ce qui stabilise les décisions inter-runs.
- Optimisation multi-objectifs (p. ex. Optuna) sur $(\text{PR-AUC}\uparrow, \text{Cost}@t^\star\downarrow, \text{ECE}\downarrow)$ avec contrainte de calibration.

3.3 Enrichissement externe (*vraies données*, neutres vis-à-vis du modèle)

Par `Region_`. Densité de population, urbanisation, trafic moyen, accidentologie locale (accidents/km, gravité), pluviométrie, jours de gel.

Par `Code_vehicule` / `Classe_vehicule_` / `Carrosserie_`. Notes de sécurité (Euro NCAP), présence/année d'ABS/ESP, ADAS, statistiques de vol, fréquences de panne.

Par `Alimentation_`. Émissions CO_2, couple/puissance normalisés, âge moyen du parc par carburant.

Par `Date_`. Saisonnalité/météo locales (pluie/neige, luminosité), jours fériés/événements, heures de pointe.

Par `CSP_` / `Situation_maritale_` / `Habitation_`. Agrégats INSEE (revenu médian, chômage, part urbaine, motorisation des ménages).

Bonus.

- **GPU CatBoost** (`task_type=GPU`) : accélération $\times 3 \sim \times 5$.
- **Contraintes de monotonie** si métier requis (p. ex. expérience $\uparrow \Rightarrow$ risque $\downarrow$) pour des extrapolations stables.
- **Ensemble léger** : moyenne pondérée des probabilités calibrées (classifieur & régressseur) validée sur set *out-of-time*.
- **Interprétabilité** : SHAP global et par segment (`Region`, `Usage`, `Carrosserie`) pour règles actionnables (pricing/contrôle).

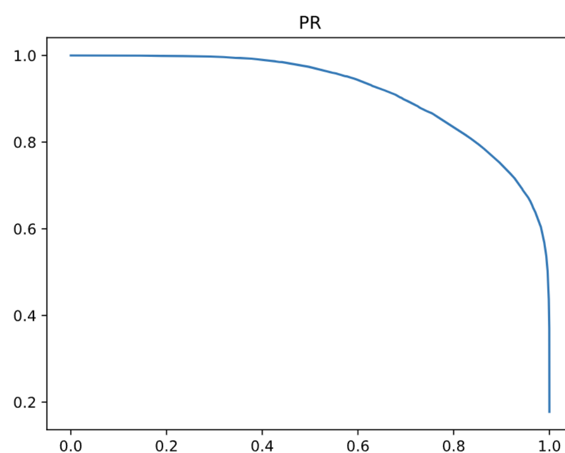
Annexe: definitions

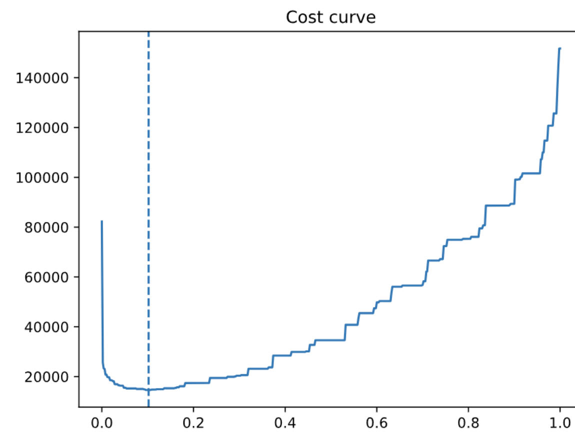
$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \qquad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$
$$\text{F1} = \frac{2\,\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \qquad \text{ROC-AUC} = \int \text{TPR}(\text{FPR})\, d\text{FPR}.$$

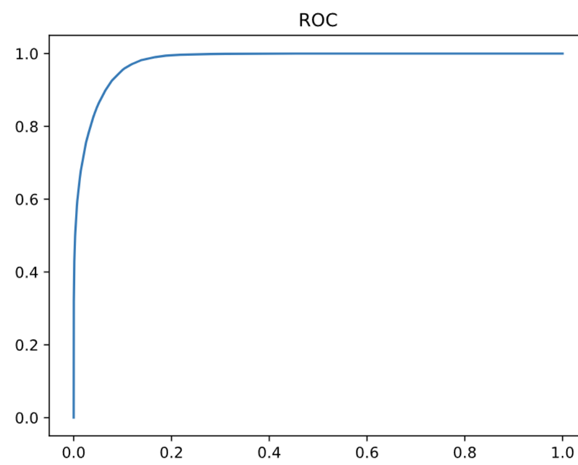
Annexe: graphiques

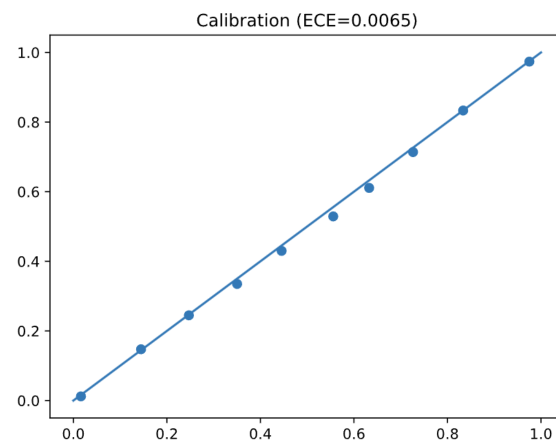


Figure 1: Schema Architecture









Predicted Category

Actual Category	no-claim	claim
no-claim	86.09	13.9
claim	1.77	98.22

Actual Category \ Predicted Category	no-claim	claim
no-claim	70.8k	11.4k
claim	314	17.4k